Team Members
Mohamed Ibrahim Seada
Fouad Mohamed Faid

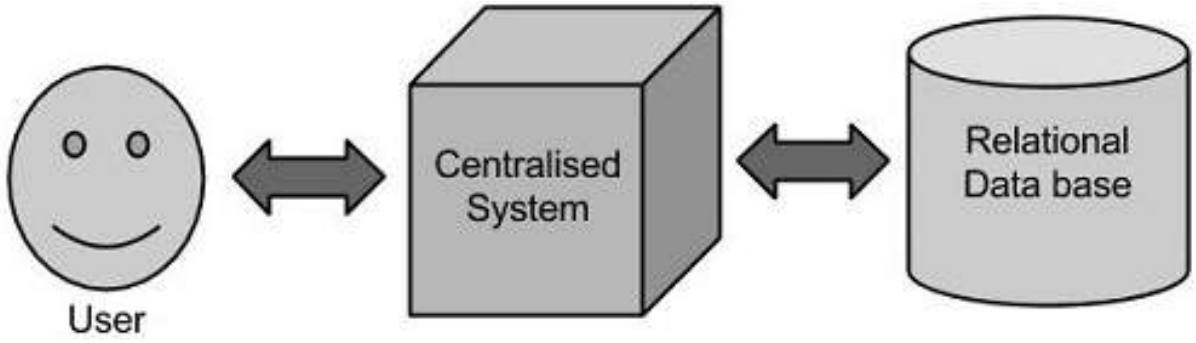
Big Data

Big Data Overview

- Due to new technologies like social networking sites, the data amount produced is growing rapidly every year.
- Big data technologies are important in providing accurate analysis, which may lead to more concrete decision-making resulting.

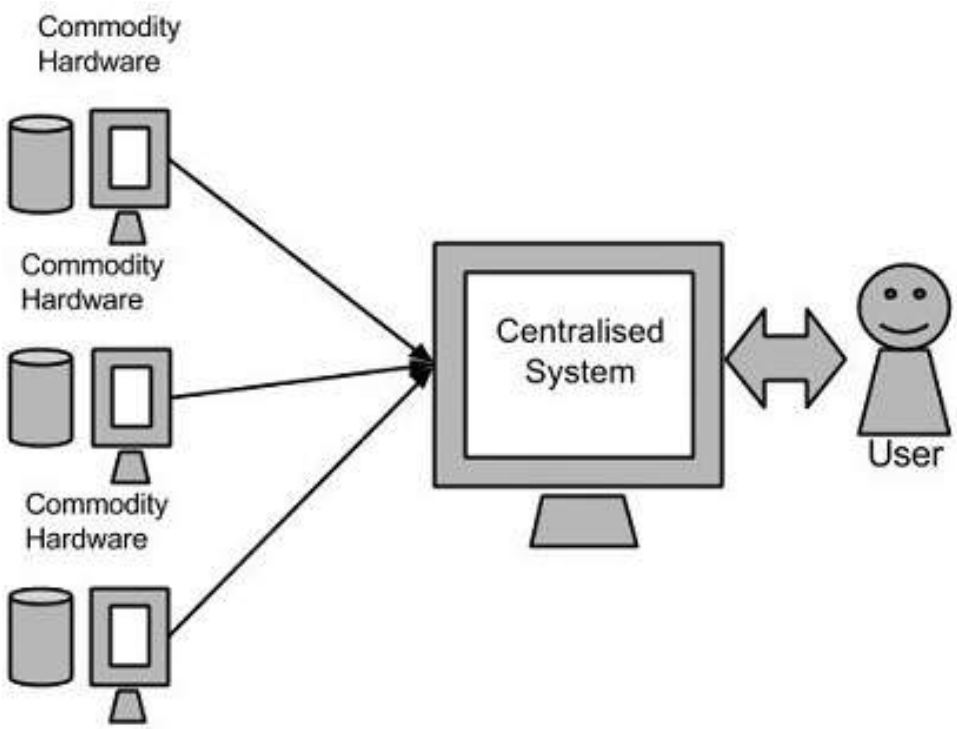
Big Data Overview

- The term "Big Data" is used to describe the collection of complex and large data sets.
- It's difficult to store and process this kind of data using traditional databases management system.



Big Data Overview

- To deal with this kind of complex data, we need to new technologies that used distributed systems.

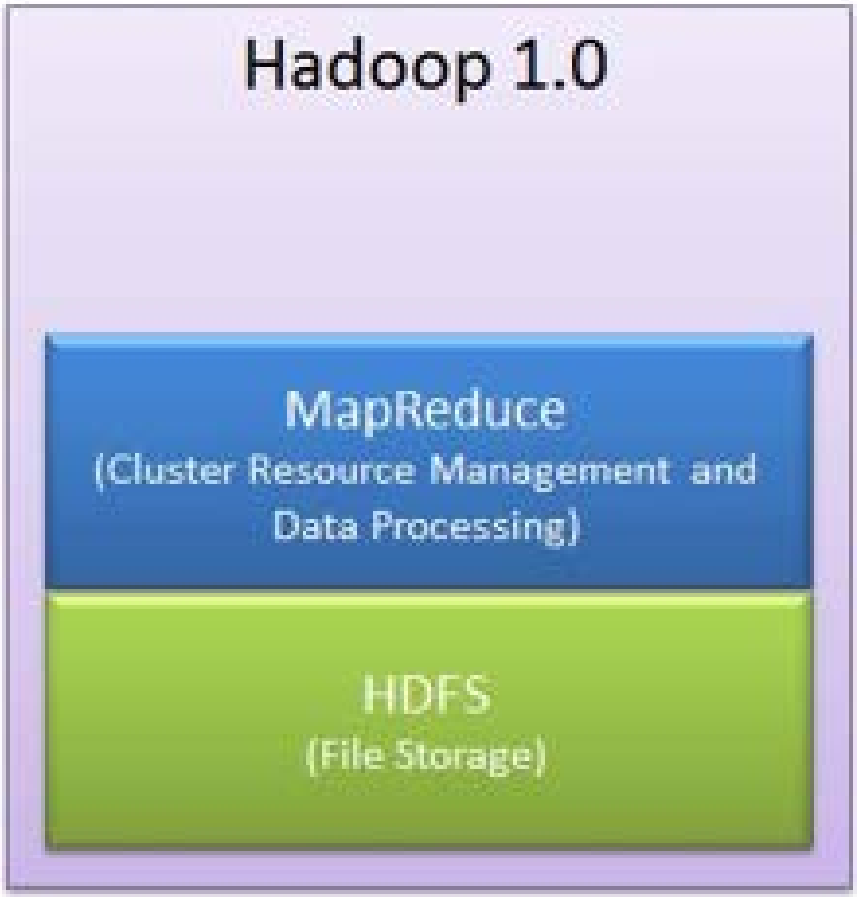


Hadoop

Hadoop Overview

- Hadoop is an Open-source Apache project that allows to store and process big data in a distributed environment.
- Hadoop was created out of Yahoo! in 2006 and It was inspired by Google's MapReduce framework.
- After years of development within the open source community, Hadoop 1.0 became publicly available in November 2012.

Architecture



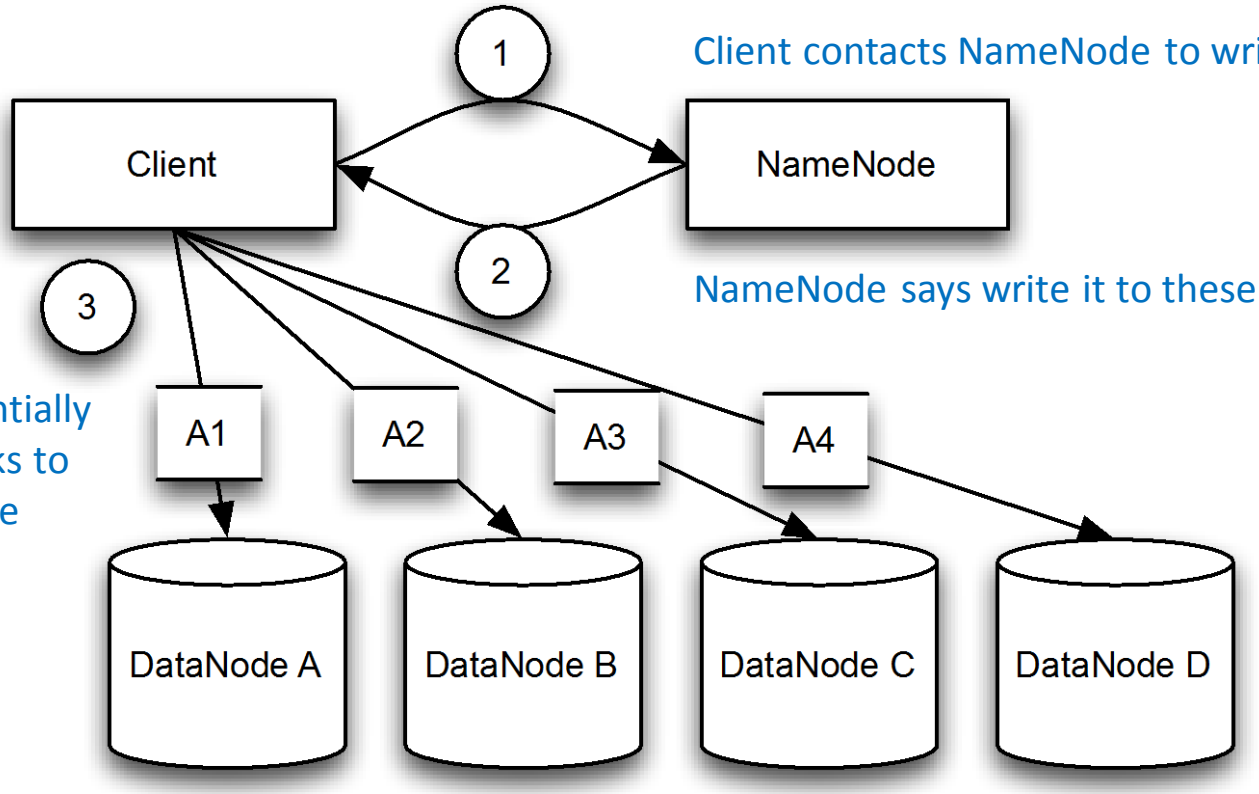
HDFS

Hadoop Distributed File System

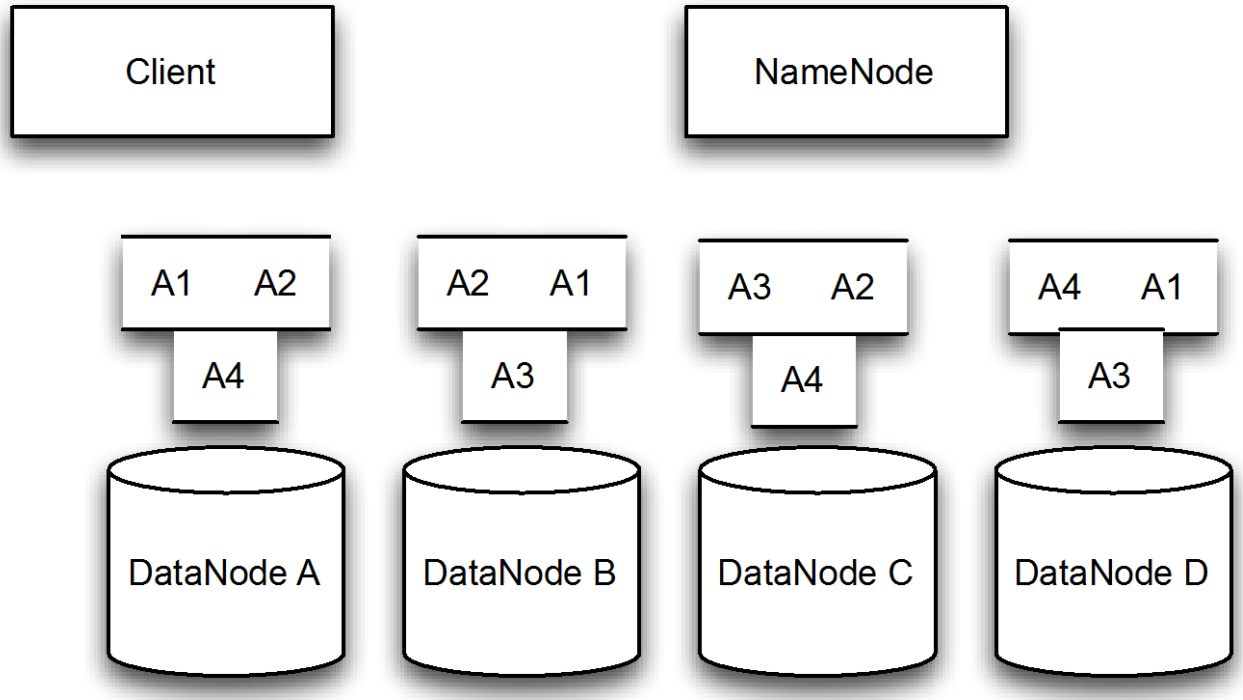
HDFS Overview

- Hierarchical UNIX-like file system for data storage.
  - It contains files, directories, permissions, users, and groups.
- Splitting of large files into blocks.
- Distribution and replication of blocks to nodes.
- Two key services
  - Master NameNode
  - Many DataNodes
- Checkpoint Node (Secondary NameNode).
  - It is not a hot backup!

How HDFS Works - Writes

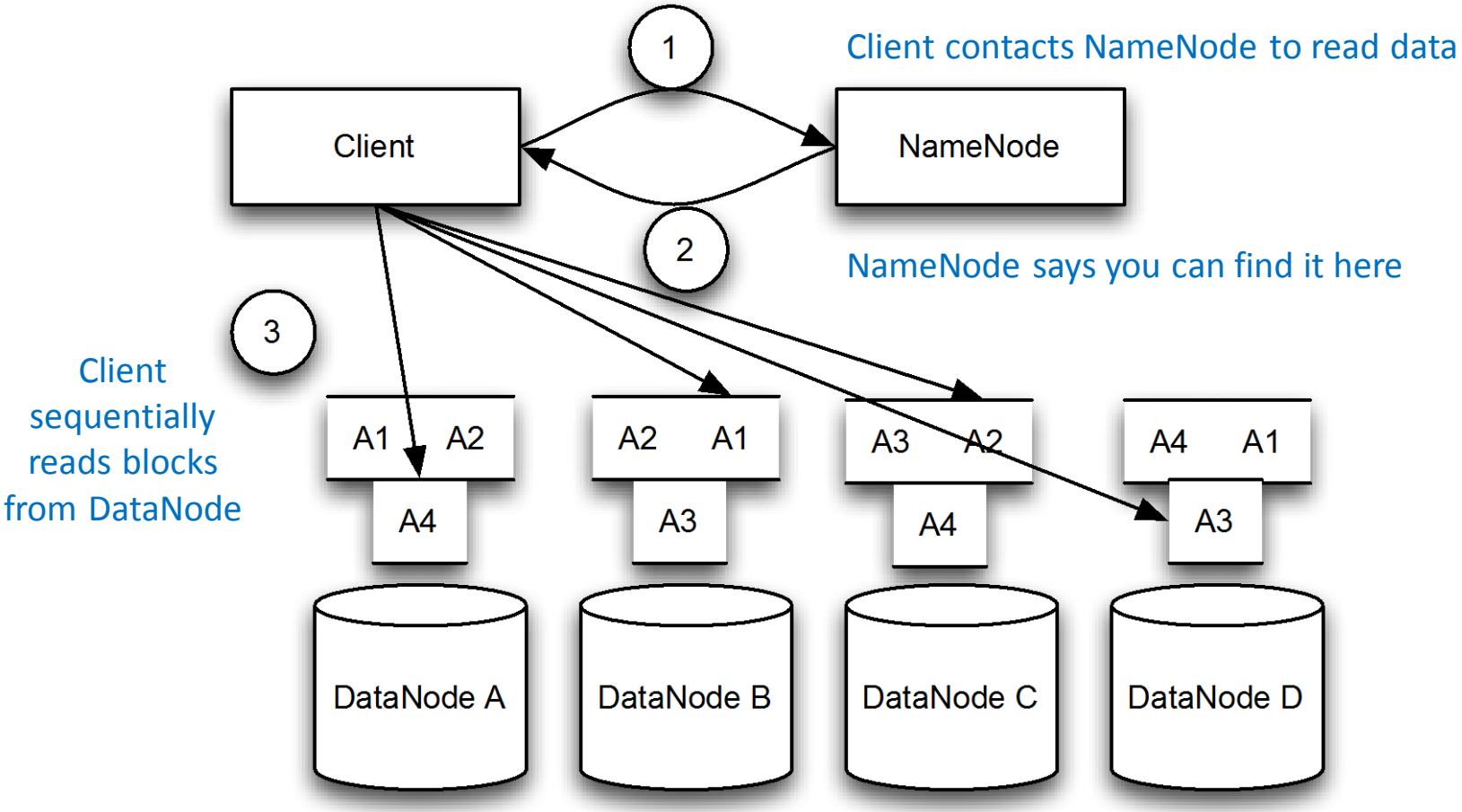


How HDFS Works - Writes

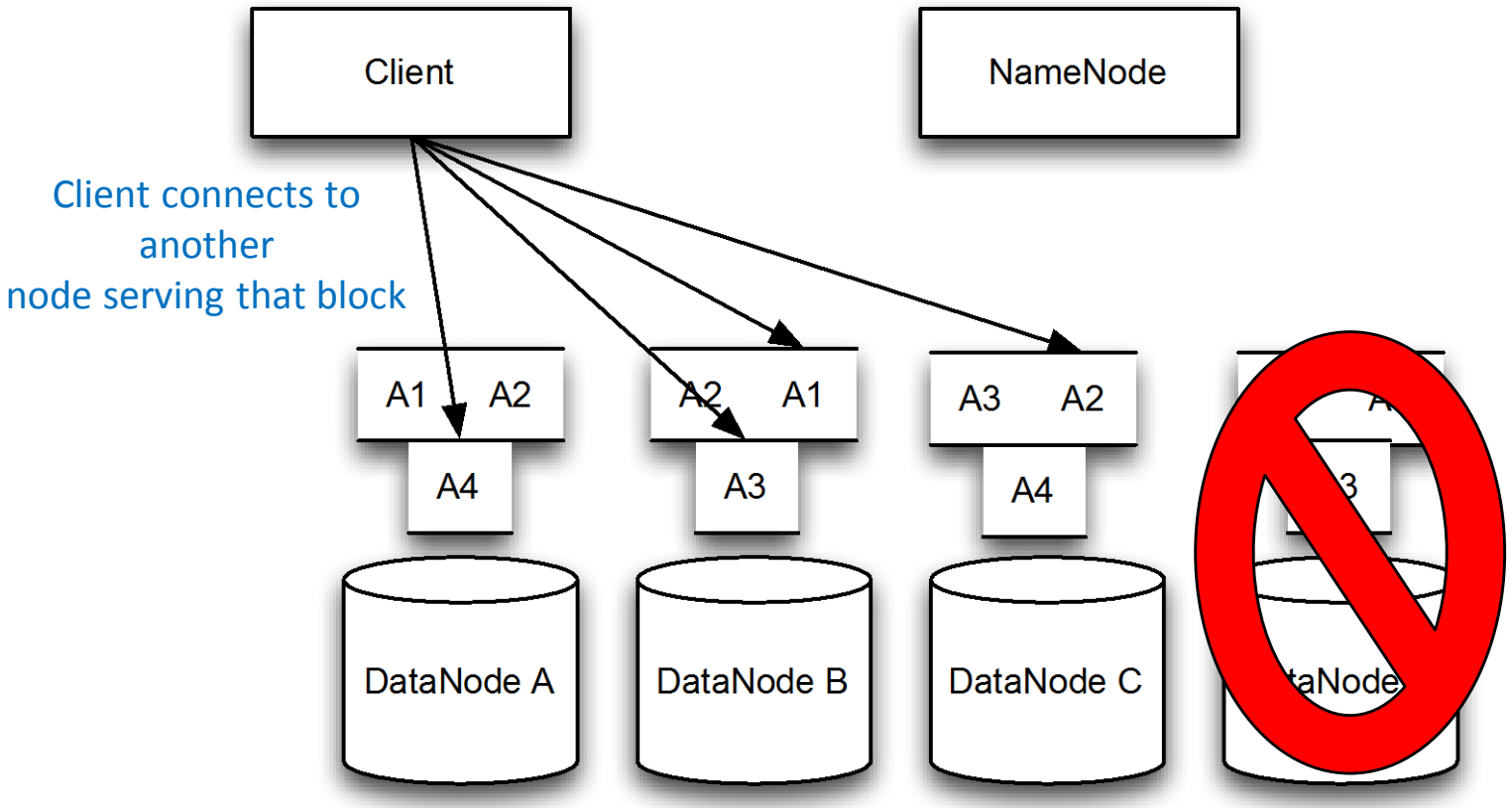


DataNodes replicate data blocks, orchestrated by the NameNode

How HDFS Works - Reads

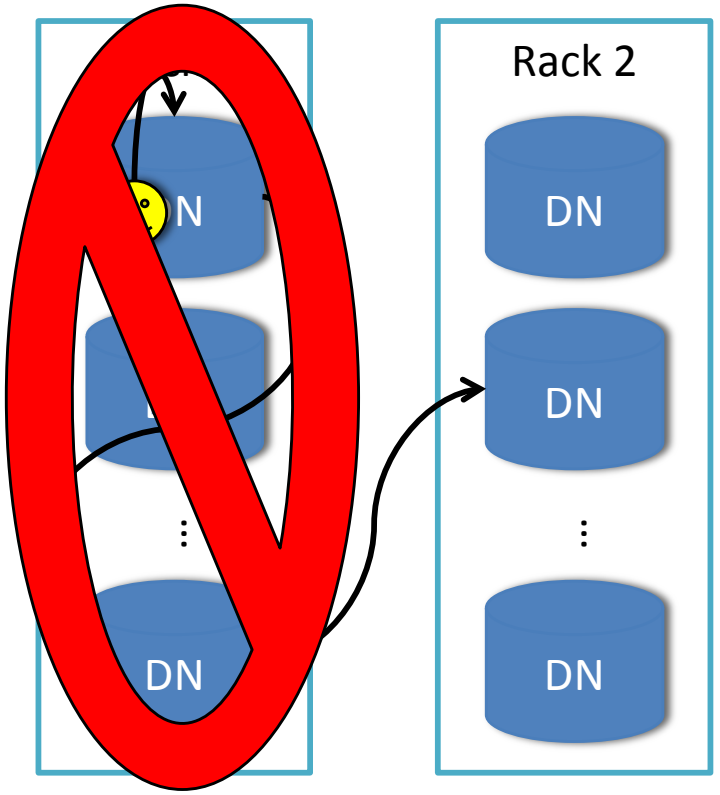


How HDFS Works - Failure



Block Replication

- Default of three replicas
- Rack-aware system
  - One block on same rack
  - One block on same rack, different host
  - One block on another rack
- Automatic re-copy by NameNode, as needed



HDFS 2.0 Features

NameNode High-Availability (HA)

- Two redundant NameNodes
  - ⇨ Primary NameNode
  - ⇨ Hot standby NameNode
- Manual or automated failover.

NameNode Federation

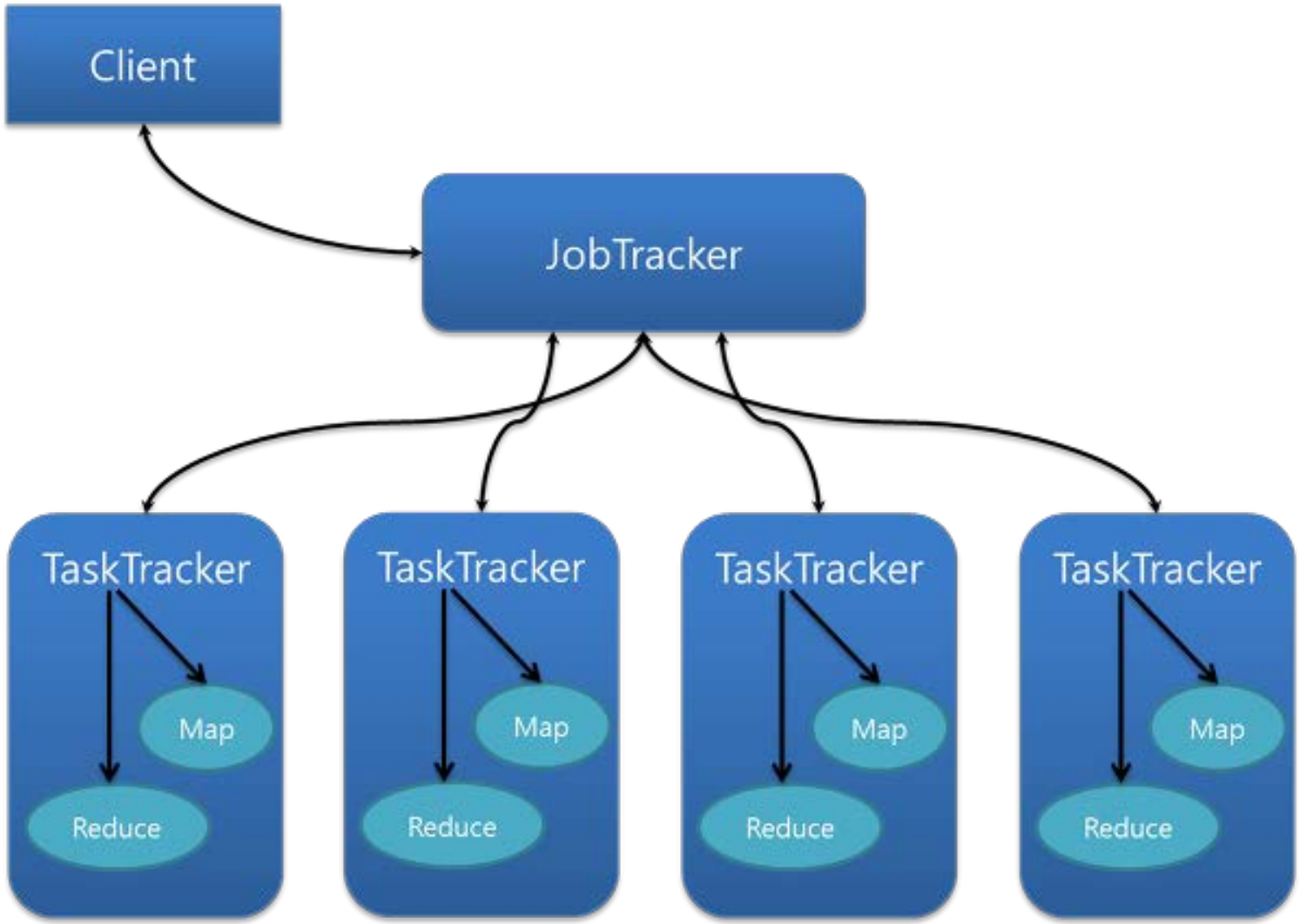
- The new storage architecture generalizes the block storage layer.
- It can be used not only by HDFS but also other storage services.

MapReduce

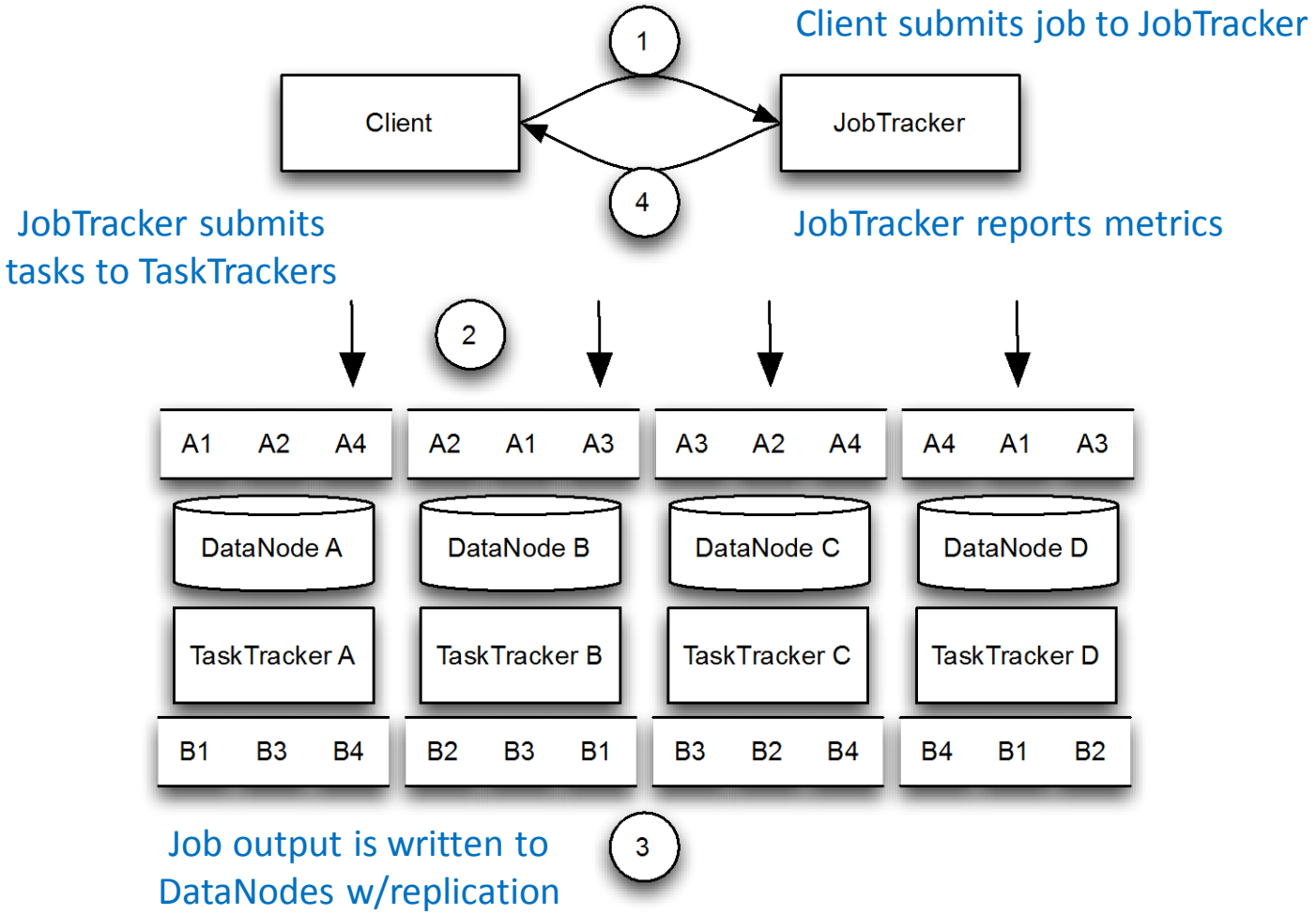
MapReduce

- Hadoop MapReduce is a software framework for easily writing applications which process big amounts of data in-parallel on large clusters.
- Hadoop Map-Reduce framework works on Master/Slave architecture.
- Map-Reduce consists of two components:
  - Task Allocation
  - Data processing

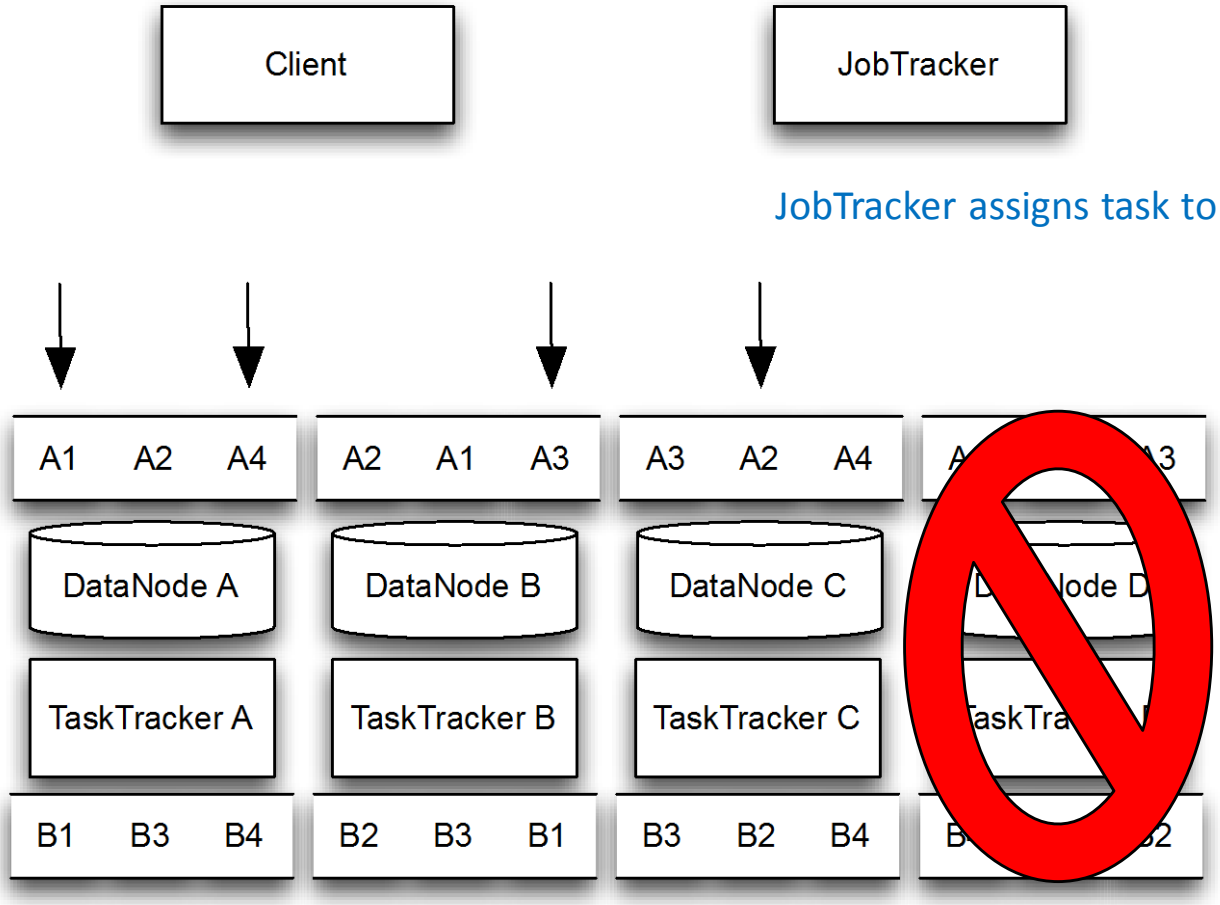
MapReduce Architecture



Task Allocation



Task Allocation - Failure



Exploiting Data Locality

- JobTracker will schedule task on a TaskTracker that is local to the block
  - 3 options!
- If TaskTracker is busy, selects TaskTracker on same rack
  - Many options!
- If still busy, chooses an available TaskTracker at random
  - Rare!

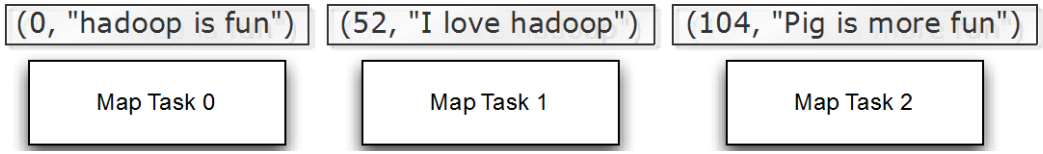
Data Processing

It consist of two key phases

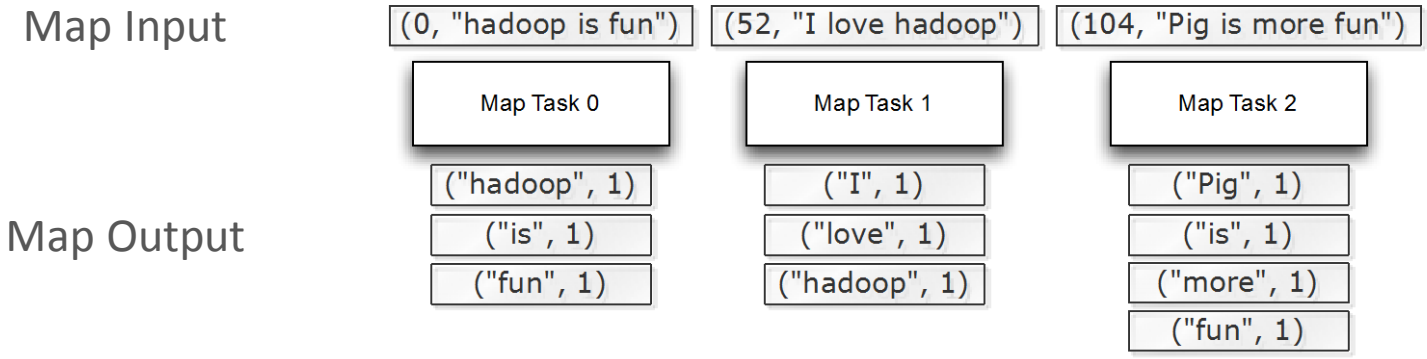
- **Map**
  - ⇨ a procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name).
- **Reduce**
  - ⇨ a procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies).

Word Count Example

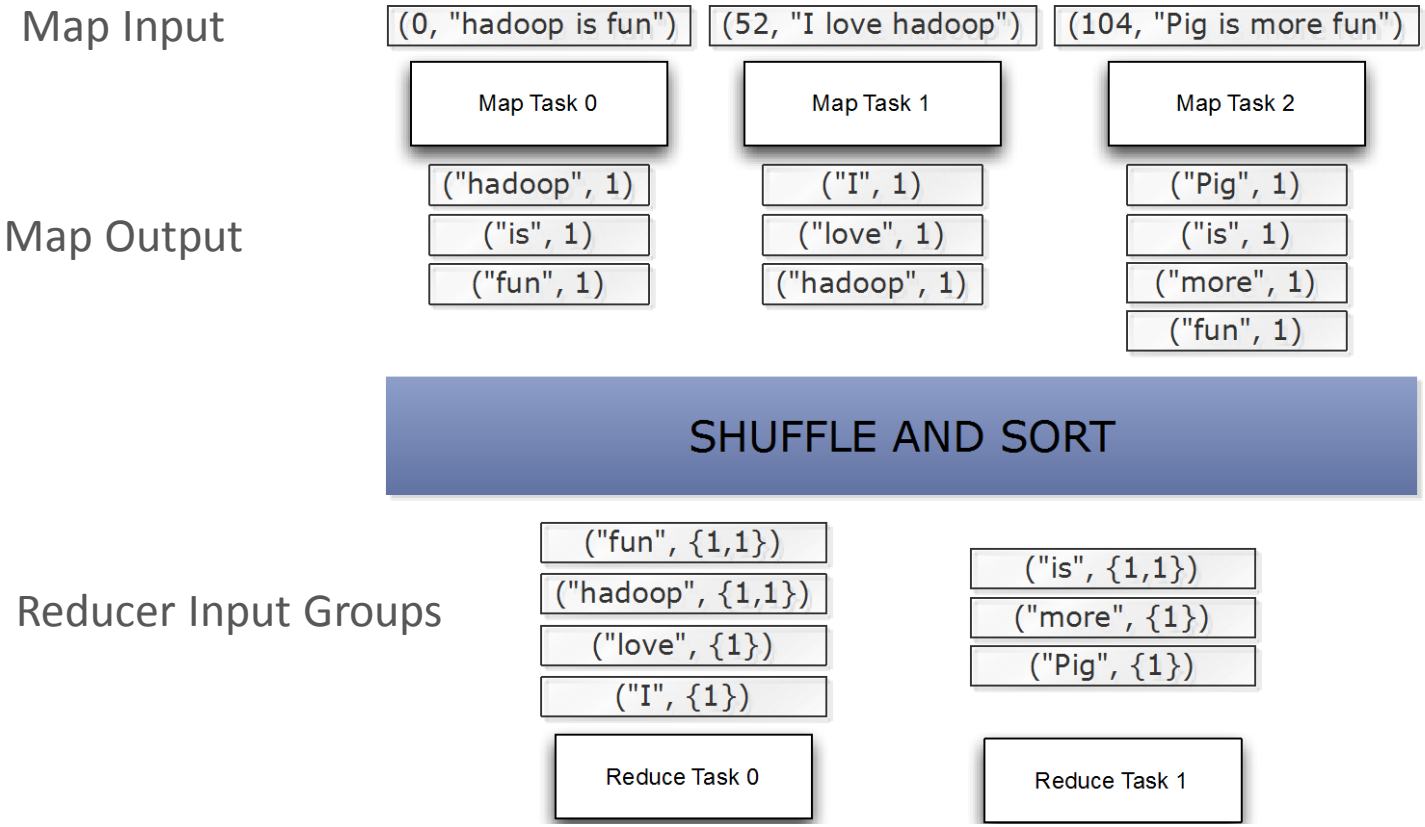
Map Input



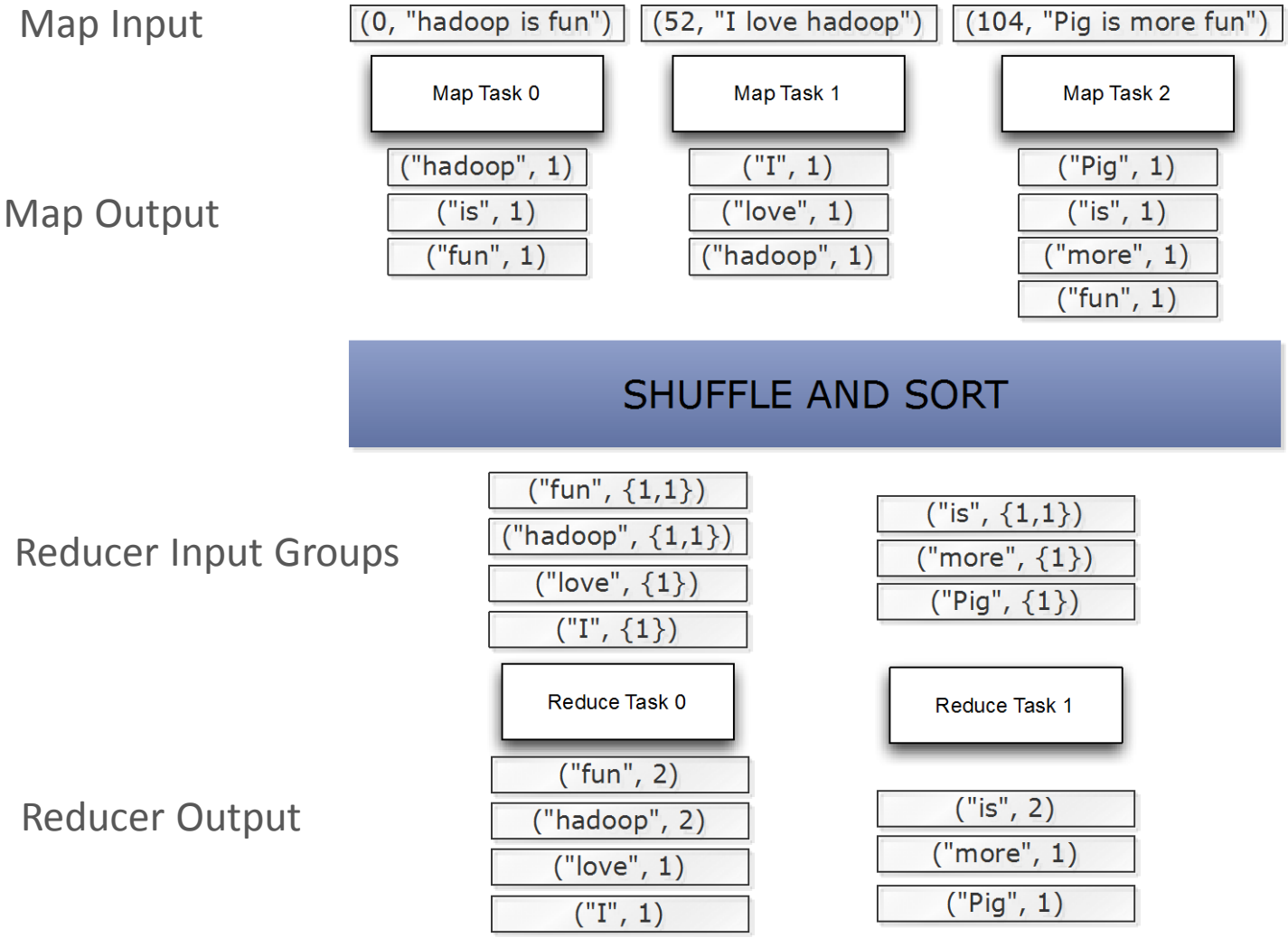
Word Count Example



Word Count Example



Word Count Example

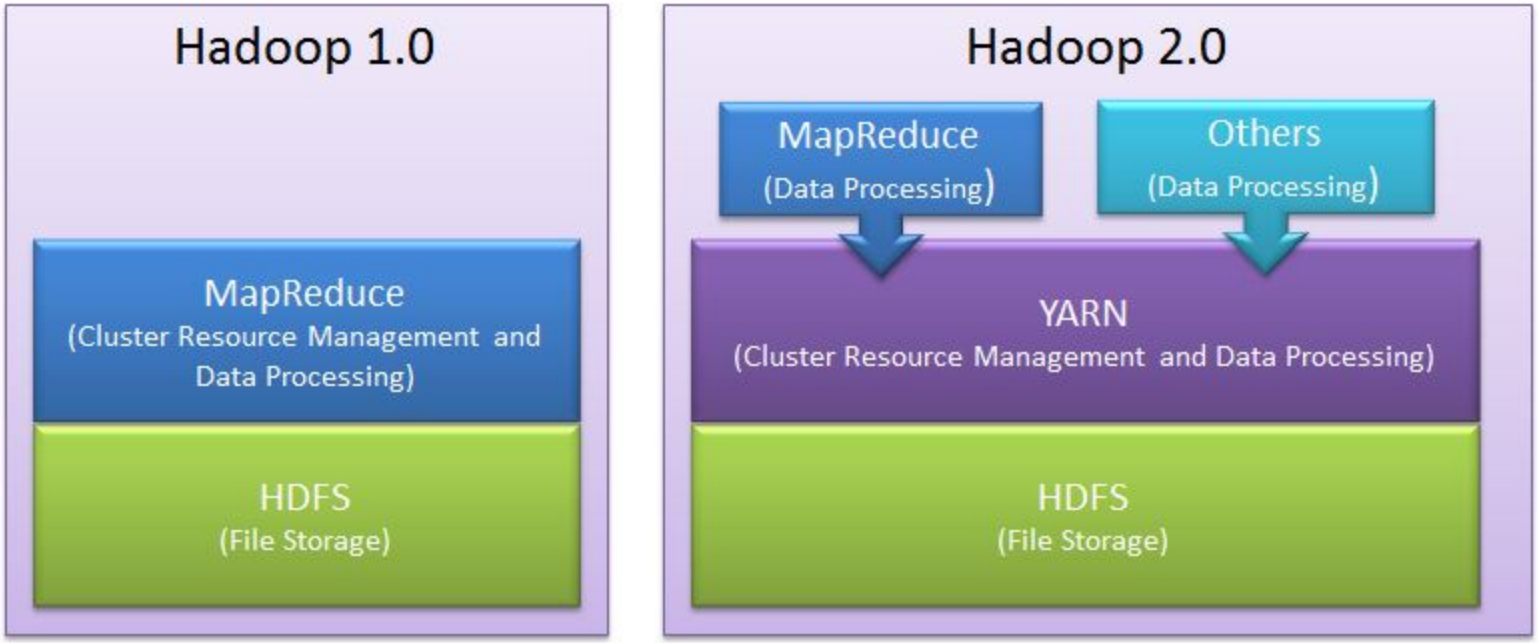


YARN

Yet Another Resource Negotiator

YARN Overview

- YARN is a new component added in Hadoop 2.0
- It is a framework which is responsible for doing Cluster Resource Management.



YARN Overview

- YARN is backward compatible.
  - The existing MapReduce job can run on Hadoop 2.0 without any change.
- YARN Splits functionality of JobTracker into
  - Resource Manager
  - Application Master
- TaskTracker becomes NodeManager

Why YARN?

Problems with this approach in Hadoop 1.0

- **It limits scalability**
  - ⇨ JobTracker runs on single machine doing several task.
  - ⇨ There are so many (DataNode) available; they are not getting used.
- **Availability Issue**
  - ⇨ In Hadoop 1.0, JobTracker is single Point of availability. This means if JobTracker fails, all jobs must restart.
- **Problem with Resource Utilization**
  - ⇨ In Hadoop 1.0, there is concept of predefined number of map slots and reduce slots for each TaskTrackers.
  - ⇨ Resource Utilization issues occur because maps slots might be 'full' while reduce slots is empty (and vice-versa).
- **Limitation in running non-MapReduce Application**

Why YARN?

Advantages of YARN

- **YARN does efficient utilization of the resource.**
  - ⇨ There are no more fixed map-reduce slots.
  - ⇨ YARN provides central resource manager.
  - ⇨ With YARN, you can now run multiple applications in Hadoop, all sharing a common resource.
- **Yarn can even run application that do not follow MapReduce model.**


THANK
YOU